Optimal Unsupervised Learning

T. L. H. Watkin
Laboratoire de Physique Statistique de L'Ecole Normale Supérieure, Paris
and (permanent address):
St. John's College
Cambridge, CB2 1TP, U.K.

J.-P. Nadal
Laboratoire de Physique Statistique[*]
Ecole Normale Supérieure
24, rue Lhomond, 75231 Paris Cedex 05 - France

Abstract

We introduce an inferential approach to unsupervised learning which allows us to define an optimal learning strategy. Applying these ideas to a simple, previously studied model, we show that it is *impossible* to detect structure in data until a critical number of examples have been presented – an effect which will be observed in all problems with certain underlying symmetries. Thereafter the advantage of optimal learning over previously studied learning algorithms depends critically upon the distribution of patterns; optimal learning may be exponentially faster. Models with more subtle correlations are harder to analyse, but in a simple limit of one such problem we calculate exactly the efficacy of an algorithm similar to some used in practice, and compare it to that of the optimal prescription.

PACS. 87.10 - PACS. 02.50 - PACS. 64.60C
Published in *J. Phys. A: Math. and Gen.*

[*]Laboratoire associé au C.N.R.S. (U.R.A. 1306), à l'E.N.S. et aux Universités Paris VI et Paris VII.

1 Introduction.

Great successes have been achieved during the past few years in applying statistical mechanics to the analysis of how neural networks learn computational tasks from examples of what must be done – so-called *supervised learning*. Starting from very simple toy models, this field has eventually led to an understanding even of problems whose complexity approaches that of reality. This topic is comprehensively reviewed by [1].

A related, but different, problem, is *unsupervised learning*. Instead of being told how the data, a set of input patterns, is to be classified (what is the desired output for each pattern), one is interested in deducing the *distribution* of the inputs. The patterns might, for example, be distributed in several clusters; one would like to learn where the clusters are, and thus recognise from which cluster a new pattern is drawn.

Practically, this may be done using an empirical algorithm [2], or a criterion (a cost function for the adaptive parameters) which characterizes the quality of the output distribution. The cost function (or the algorithm) is chosen according to the particular task and the specific constraints one is interested in. One example is the use of information theoretic criteria in modeling early stages of information processing in the brain [3, 4, 5, 6]. In this approach the emphasis is put on the representation (the coding) of the data. In the simplest cases, one ends up with a principal component analysis of the data [7, 8, 4]. Other examples are the topological mapping algorithm of Kohonen [2] and related algorithms based on cost functions [9], and Kohonen's Learning Vector Quantization algorithm [2, 10]. These examples are related to what is known as *clustering* in data analysis: one hopes to infer *structure* in the data.

In this paper we are interested in the use of statistical mechanics for studying clustering algorithms. By analogy with the theory of supervised learning [1], one can identify several ways in which statistical mechanics might prove useful:

- it can suggest learning strategies based on Monte-Carlo algorithms - simulated annealing has already been proposed [11];
- it could predict the success of given learning algorithms which minimise a cost function - a recent paper has applied this idea to a simple model of an unsupervised learning task [12];
- and it could quantify the information which can be extracted in principle from the data, and thus describe optimal learning algorithms.

Here we concentrate on the last aspect. We present a general, inferential formalism for unsupervised learning and explain what optimal learning of an unsupervised task would mean. To make the formalism a little more transparent, we compare optimal learning to the best results obtained by [12]. We then study a model in which the underlying structure of patterns is more complicated, and exactly compare optimal learning to an algorithm which resembles those already used in reality.

2 Unsupervised Learning and Optimal Learning

2.1 Inferential Formalism

Let us suppose that we have p N-vector patterns, $\{\boldsymbol{\xi}^\mu\}$, $\mu = 1, \ldots, p$, drawn from an unknown distribution $\mathcal{P}(\boldsymbol{\xi})$, which we would like to infer. Our data is called the *training set*. Typically $P(\boldsymbol{\xi})$ might be such that patterns are correlated with one or more of K *prototype* N-vectors, $\{\mathbf{B}^l\}$, $l = 1, \ldots, K$. The prototypes are normalised such that $\mathbf{B}^l \cdot \mathbf{B}^l = 1$ for all l (note that this paper consistently takes the definition of the scalar product of two N vectors as $\mathbf{a} \cdot \mathbf{b} \equiv \frac{1}{N} \sum_i a_i b_i$). If $K \ll p$, then the training set will be arranged in clusters around the prototypes. We would like to know the $\mathcal{B} \equiv \{\mathbf{B}^l\}$ and the exact form of the correlations.

To frame any model of $\mathcal{P}(\boldsymbol{\xi})$, we must make some guess about its complexity. We might, for example, hypothesise about the value of K, and assume that each cluster is spherically symmetric about one of the prototypes. The set of all models of this form is our *hypothesis space*. We would use the training set to fit the remaining parameters in the model. If the model is too simple then enough training data will tell us so, and we would generalise it to a more complex one.

Given the form of the model, how exactly should we fit the parameters? A vast literature exists on practical strategies for so doing, and many algorithms have turned out to be useful. A better understanding of the problem would be desirable, if the algorithms are going to be systematically improved, and might also tell us about the methods' limitations. A start in this direction has been made by Biehl and Mietzner [12] who introduced a simple, solvable unsupervised task, and analysed learning it with two intuitively reasonable algorithms.

To define an optimal way to guess the $\{\mathbf{B}^l\}$, we present an inferential framework. Before we have any data about the model, the $\mathcal{B} \equiv \{\mathbf{B}^l\}$ are totally unknown, so our ignorance is expressed in a uniform *prior probability distribution*: $P_0(\{\mathbf{B}^l\}) = \prod_l P_0(\mathbf{B}^l)$, where $P_0(\mathbf{B}^l)$ is a constant for all correctly normalised $\mathbf{B}^l$. Let the variable σ^μ label which cluster $\boldsymbol{\xi}^\mu$ is drawn from, $\sigma^\mu \in 1, \ldots, K$. We will use $q_0(\sigma^\mu)$ to be the prior probability that the μth question is from σ^μth cluster. Our simplest models will assume that these values are known, and in fact that $q_0(\sigma^\mu) = 1/K$ for all σ^μ.

If we knew the $\mathcal{B}$, and the form of the clusters, we could write the probability that a question $\boldsymbol{\xi}^\mu$ is generated as

$$P(\boldsymbol{\xi}^\mu/\{\mathbf{B}^l\}) = \underset{\{\sigma^\mu\}}{Tr}\, P(\boldsymbol{\xi}^\mu/\{\mathbf{B}^l\}, \sigma^\mu) q_0(\sigma^\mu) \tag{1}$$

where $P(\boldsymbol{\xi}^\mu/\{\mathbf{B}^l\}, \sigma^\mu) = P(\boldsymbol{\xi}^\mu/\mathbf{B}^{l=\sigma^\mu})$ characterizes the σ^μth cluster.

The posterior probability for $\{\mathbf{B}^l\}$ is given by *Bayes rule*:

$$P(\{\mathbf{B}^l\}/\{\boldsymbol{\xi}^\mu\}) = \frac{\prod_\mu P(\boldsymbol{\xi}^\mu/\{\mathbf{B}^l\}) P_0(\{\mathbf{B}^l\})}{Z}, \tag{2}$$

where $Z = Z(\boldsymbol{\xi}^\mu)$ is the normalisation constant

$$\begin{aligned} Z &\equiv \prod_l \int \mathrm{d}\mathbf{B}^l \prod_\mu P(\boldsymbol{\xi}^\mu/\{\mathbf{B}^l\}) P_0(\{\mathbf{B}^l\}) \\ &= \mathop{Tr}_{\{\sigma^\mu\}} \prod_l \int \mathrm{d}\mathbf{B}^l \prod_\mu P(\boldsymbol{\xi}^\mu/\{\mathbf{B}^l\}, \sigma^\mu)\, q_0(\sigma^\mu) P_0(\{\mathbf{B}^l\}). \end{aligned} \tag{3}$$

$P(\{\mathbf{B}^l\}/\{\boldsymbol{\xi}^\mu\})$ is a useful quantity. For example, while it is easy to measure the success of algorithms for supervised learning (because as the training set grows we can continually compare how well our algorithm predicts the answers to the new questions with the given right answers), it is harder to compare the success of algorithms for unsupervised learning. We do not know in really which clusters new examples "really" come from, so how can we measure how well the unsupervised learning of the clusters has been? The inferential approach lets us do this, at least in principle. If we knew the true $\mathcal{P}(\boldsymbol{\xi})$, then we could define the *quality* of the hypothesis $\{\tilde{\mathbf{B}}^l\}$ as some function $\mathcal{Q}(\{\tilde{\mathbf{B}}^l\}, \{\mathbf{B}^l\})$ (for example $\mathcal{Q} = \sum_l \tilde{\mathbf{B}}^l \cdot \mathcal{B}^l$). Since $\mathcal{B}$ is not known, the expected quality of $\{\tilde{\mathbf{B}}^l\}$ is $\left\langle \mathcal{Q}(\{\tilde{\mathbf{B}}^l\}, \{\mathbf{B}^l\}) \right\rangle_{\{\mathbf{B}^l\}}$, where the average is over choices of $\{\mathbf{B}^l\}$ from (2). Indeed, in a high dimensional space such a quantity may be self-averaging, which means that the true (unknown) value of $\mathcal{Q}(\{\tilde{\mathbf{B}}^l\}, \mathcal{B})$ is very close to $\left\langle \mathcal{Q}(\{\tilde{\mathbf{B}}^l\}, \{\mathbf{B}^l\}) \right\rangle_{\{\mathbf{B}^l\}}$.

The posterior probability of $\{\mathbf{B}^l\}$ also allows us to invent learning algorithms. An obvious way of chosing a hypothesis $\{\tilde{\mathbf{B}}^l\}$, for example, would just be a sample from (2); by analogy with supervised theory, we call this Gibbs learning.

A more sophisticated strategy would be to choose to be the maximum of Eqn. (2), the *principle of maximum a postiori probability* (MAPP). Unfortunately, the maximum of many distributions is a long way away from a large proportion of the possible samples, so MAPP may give a very poor approximation to the right classification.

However, only a paragraph ago, we were able to define the expected quality of $\{\tilde{\mathbf{B}}^l\}$, relative to a quality measure $\mathcal{Q}$. The $\{\tilde{\mathbf{B}}^l\}$ maximising $\left\langle \mathcal{Q}(\{\tilde{\mathbf{B}}^l\}, \{\mathbf{B}^l\}) \right\rangle_{\{\mathbf{B}^l\}}$ is optimal, in the sense of maximum quality. Generating this $\{\tilde{\mathbf{B}}^l\}$ we call *optimal learning.*

Note that Z includes the average over all possibilities: of the $\{\mathbf{B}^l\}$ and the correct labeling of the patterns in the training set $\{\sigma^\mu\}$. It may equally well be used to motivate ways to guess from which clusters the patterns in the training set were drawn, since the posterior probability of $\{\sigma^\mu\}$ is

$$P(\{\sigma^\mu\}/\{\boldsymbol{\xi}^\mu\}) = \frac{\int \mathrm{d}\mathbf{B} \prod_\mu P(\boldsymbol{\xi}^\mu/\{\mathbf{B}^l\}, \sigma^\mu) q_0(\sigma^\mu) P_0(\{\mathbf{B}^l\})}{Z}. \tag{4}$$

By analogy with the nomenclature above, a hypothesis $\{\tilde{\sigma}^\mu\}$ taken from (4) could be called a Gibbs prediction. If we define a measure $\tilde{\mathcal{Q}}(\{\tilde{\sigma}^\mu\}, \{\sigma^\mu\})$ of the quality of $\{\tilde{\sigma}^\mu\}$ given $\{\sigma^\mu\}$, then the optimal way of choosing $\{\tilde{\sigma}^\mu\}$ would be so as to maximise the expectation value of $\left\langle \tilde{\mathcal{Q}}(\{\tilde{\sigma}^\mu\}, \{\sigma^\mu\}) \right\rangle_{\{\sigma^\mu\}}$, where $\langle \ldots \rangle_{\{\sigma^\mu\}}$ means the average over $\{\sigma^\mu\}$ from (4). A natural choice for $\tilde{\mathcal{Q}}$ is $\frac{1}{p} \sum_\mu \delta(\tilde{\sigma}^\mu, \sigma^\mu)$. In this case, the best choice for each $\tilde{\sigma}^\mu$ is such as to equal the σ^μ which maximises

$$P(\sigma^\mu/\{\boldsymbol{\xi}^\mu\}) = \mathop{Tr}_{\{\sigma^{\mu'}\}} P(\{\sigma^{\mu'}\}/\{\boldsymbol{\xi}^\mu\}), \tag{5}$$

where the trace is over $\mu' \neq \mu$. An example is given in section 3.

2.2 Typical behaviour

Quantities such as $\left\langle \mathcal{Q}(\{\tilde{\mathbf{B}}^l\}, \{\mathbf{B}^l\}) \right\rangle_{\{\mathbf{B}^l\}}$ may be calculated from derivatives of $\ln Z$, by introducing appropriate field terms. In statistical mechanics, we are interested in calculating not the properties a single data set, but of *typical* data. In fact for large N we expect

$$\mathcal{L}(\boldsymbol{\xi}^\mu) \equiv \ln Z(\boldsymbol{\xi}^\mu) \tag{6}$$

to be a self-averaging quantity, that is

$$lim_{N\to\infty}\mathcal{L}(\boldsymbol{\xi}^\mu)/N = lim_{N\to\infty} \left\langle \mathcal{L}(\boldsymbol{\xi}^\mu) \right\rangle_\xi /N \tag{7}$$

where $\langle \ldots \rangle_\xi$ indicates the average over realisations of $\{\boldsymbol{\xi}^\mu\}$.

The average of $\mathcal{L}$ is equal to the logarithm of the model distribution of the patterns $\boldsymbol{\xi}^\mu$, averaged over the true distribution $\mathcal{P}(\boldsymbol{\xi}^\mu)$. Hence, up to a constant it is equal to minus the *Kulbback-Leibler distance* $\mathcal{D}$ of the estimated distribution Z relative to the true distribution $\mathcal{P}$:

$$\mathcal{D} = \int d\boldsymbol{\xi}^\mu \mathcal{P} \ln \frac{\mathcal{P}}{\mathcal{Z}} \tag{8}$$

In statistical inference [13, 14] this quantity is known to be a relevant measure of the discrepency beween the two distributions. Maximizing $\ln Z$ over the parameters that define the B distribution is thus equivalent to minimizing the distance $\mathcal{D}$.

For a simple enough model, the average (7) may be performed in a now standard way, using the method of replicas [1]. In this way one will get the average (typical value) of $\mathcal{Q}$, whereas we are interested in finding the optimal $\{\tilde{\mathbf{B}}^l\}$ for a given training set. However, as we show now, there is however a particular case, which we expect to be generic, in which both the analytical study and the algorithmic implementation can be performed.

2.3 A favorable case

In supervised learning a certain choice for the analogue of the $\mathcal{Q}$ factor was very natural (so that optimal supervised learning could be easily defined [15]). In unsupervised learning it is not so clear. Nevertheless, it may well be that the optimal $\{\tilde{\mathbf{B}}^l\}$ is only weakly dependent on $\mathcal{Q}$ (as in the examples given in the following sections). For example, by analogy with supervised learning [1], we expect often to find that if $\{\tilde{\mathbf{B}}^l\}^\alpha$ and $\{\tilde{\mathbf{B}}^l\}^\beta$ are two hypotheses about $\mathcal{B}$ drawn from (2), then for N large the value of $\tilde{\mathbf{B}}^{l,\alpha} \cdot \tilde{\mathbf{B}}^{l,\beta}$ is distributed with a very sharp peak (a phenomenon known as *replica symmetry*). If this is true, and if $\mathcal{Q}$ may be written as $\sum_l f_l(\tilde{\mathbf{B}}^l \cdot \mathbf{B}^l)$ for some set of smooth, increasing functions $\{f_l(x)\}$, then Appendix 1 shows that the optimal $\{\tilde{\mathbf{B}}^l\}$ is *always* $\{\tilde{\mathbf{B}}^l\}^{opt}$ given by $\tilde{\mathbf{B}}^{l,opt} = \frac{1}{\gamma}\left\langle \mathbf{B}^l \right\rangle_{\{\mathbf{B}^l\}}$, where γ is a normalisation factor. For $K > 1$ and certain sorts of clusters, it is possible that there is a natural choice for $\mathcal{Q}$ which is not separable in this way. Note that this does not undermine the theory: once $\mathcal{Q}$ is chosen, $\{\tilde{\mathbf{B}}^l\}^{opt}$ is unambiguously defined.

2.4 Neural networks

So far no mention has been made of neural networks. Unlike supervised learning, for which, given the historical development of the field, it was natural to consider a neural network at once, unsupervised learning of the $\{\mathbf{B}^l\}$ themselves seems to make sense. Of course, a closely related task is the construction of a network which could classify new patterns by some arbitrary coding according to which cluster they came from. Once a measure D of the performance of the network is chosen, the optimal way of generating $\mathcal{N}$ so as to maximise this measure is $\left\langle D(\mathcal{N}, \{\mathbf{B}^l\}) \right\rangle_{\{\mathbf{B}^l\}}$. When the data is clustered, a natural quality measure seems to us to be how well the network detects the clustering. For example, if $S \equiv \mathcal{N}(\boldsymbol{\xi})$ may take K possible values, then one can define an arbitrary mapping, $\mathcal{C}$, of the set of clusters to the set of network outputs, $\mathcal{C} : \{\sigma\} \to \{\mathcal{S}\}$. Then we can define the quality D of network $\mathcal{N}$ as the maximum over $\mathcal{C}$ of the correlation between $\mathcal{C}(\sigma)$ and S. That is,

$$D(\mathcal{N}, \{\mathbf{B}^l\}) \equiv \max_{\mathcal{C}} \mathop{Tr}_{\sigma} \int \delta(\mathcal{N}(\boldsymbol{\xi}), \mathcal{C}(\sigma)) P(\boldsymbol{\xi}/\{\mathbf{B}^l\}) P(\sigma)\, d\boldsymbol{\xi}. \quad (9)$$

A simple example of this sort of measure is given in the next section.

2.5 Summary

Section 2 has explained that in order to learn $P(\boldsymbol{\xi})$, a model must be introduced. Once the model class has been chosen, there is an unambiguous way to choose the *best* model from the class, relative to a measure of quality. This strategy we call optimal learning, because it is guaranteed to maximise the expectation quality of the network it produces. The concept may be extended to optimal guessing of which clusters the training data came from and to the optimal way of producing a network to learn the task. The rest of this paper will compare the success of optimal learning with that of other algorithms, using techniques from statistical mechanics.

3 A Uni-directional model

3.1 The model

Recently, Biehl and Mietzner [12] considered a learning model in which the overlap of each N–vector pattern $\boldsymbol{\xi}$ with a single prototype, $\mathbf{B}$, is distributed as a Gaussian with standard deviation 1 about either ρ or $-\rho$. That is

$$P(\boldsymbol{\xi} \cdot \mathbf{B}\sqrt{N} = x) = \frac{1}{2} \sum_{\sigma=\pm 1} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-\sigma\rho)^2}{2}) \tag{10}$$

All patterns are taken to be normalised such that $\boldsymbol{\xi} \cdot \boldsymbol{\xi} = 1$.

Two explicit distributions of $\boldsymbol{\xi}$ spring to mind. The first, which we call *distribution 1*, is such that the component of $\boldsymbol{\xi}$ perpendicular to $\mathbf{B}$ is otherwise completely random in direction. This may be written

$$P(\boldsymbol{\xi}/\mathbf{B}) = \frac{1}{2} \sum_{\sigma=\pm 1} (2\pi)^{-\frac{N}{2}} \exp\{-\rho\sigma\boldsymbol{\xi} \cdot \mathbf{B}\sqrt{N} - \frac{\boldsymbol{\xi} \cdot \boldsymbol{\xi}}{2}\}. \tag{11}$$

An alternative, *distribution 2*, was proposed by [12]. It is such that components of patterns, ξ_i, have the same magnitude as B_i, but a sign which is only weakly correlated. Thus,

$$P(\boldsymbol{\xi}/\mathbf{B}) = \frac{1}{2} \sum_{\sigma=\pm 1} \prod_i \sum_{\sigma_i=\pm 1} \frac{1+\sigma_i\sigma\rho}{2} \delta(\xi_i - \sigma_i\sigma B_i). \tag{12}$$

In both cases, each pattern provides of the order of one bit of information about $\mathbf{B}$, so we expect that in the limit of $N \to \infty$ the number of examples we will need scales as $p = \alpha N$, for α of order 1.

Biehl and Mietzner [12] first analysed learning $\mathbf{B}$ from a training set by an analogue of the well–known "maximum stability algorithm" for supervised learning [17], but they found that a better approximation to $\mathbf{B}$ is always $\mathbf{J}^{mv}$, which is defined as the minimum of

$$E^{mv}(\mathbf{J}) = -\sum_{\mu} (\mathbf{J} \cdot \boldsymbol{\xi}^{\mu}\sqrt{N})^2. \tag{13}$$

Constucting this $\mathbf{J}^{mv}$ is called the "maximal variance algorithm", and exact numerical methods are well-known. [12] found that the maximal variance algorithm gives exactly the same performance for both distributions 1 and 2: if α is less than a critical value, $\alpha_c = 1/\rho^4$, then $\mathbf{J}^{mv}$ has zero overlap with $\mathbf{B}$ (no learning has occurred), but as α rises above this value $\mathbf{J}^{mv}$ smoothly approaches $\mathbf{B}$. As $\alpha \to \infty$, the value of $R^{mv} \equiv \frac{1}{N}\mathbf{J}^{mv} \cdot \mathbf{B}$ tends to 1 as $1 - R^{mv} = (1/\rho^2 + 1/\rho^4)/\alpha + O(1/\alpha^2)$.

3.2 Optimal learning for distribution 1

Let us now study optimal learning. We will sketch the analysis here and present it in more detail in Appendix 2. If the patterns are from distribution 1, then, as Appendix 2 shows, $< \ln Z >$ may be expressed as

$$\langle \ln \rangle Z = \min_{R^G, \tilde{R}} \left[\alpha G_1(R^G) + G_2(R^G, \tilde{R}) \right], \tag{14}$$

where R^G and $\tilde{R}$ are two order parameters and G_1 and G_2 are simple algebraic functions. The first term represents an "energy", and is minimised for high R^G; the rest is an "entropy", and is minimised for R^G small. The interpretation of R^G is $\tilde{\mathbf{B}}^\alpha \cdot \mathbf{B}$, where $\tilde{\mathbf{B}}^\alpha$ is a vector selected at random from (2). R^G is therefore the success of the Gibbs algorithm. Another interesting parameter is $q \equiv \tilde{\mathbf{B}}^\alpha \cdot \tilde{\mathbf{B}}^\beta$, the overlap of two samples of (2). Appendix 2 demonstrates that that $R^G = q$ (in fact, using the definitions in Section 2, it can be shown in general that the true rule is a typical sample of (2), and the result that $R^G = q$ follows at once).

Explicitly, R^G and $\tilde{R}$ are the solutions of the equations

$$R^G = \tilde{R}\left(1 - R^G\right) \tag{15}$$

and

$$\tilde{R} = \rho^2\alpha \int Dz \tanh\left(R^G\rho^2 + \rho\sqrt{R^G}z\right), \tag{16}$$

where $Dz \equiv dz \exp\left(-z^2/2\right)/\sqrt{2\pi}$, and, as in the rest of this paper, the limits of the integral are $-\infty$ to $+\infty$ unless otherwise stated.

Due to the spherical symmetry of the problem, all acceptable measures of the quality of a hypothesis $\tilde{\mathbf{B}}$ about $\mathbf{B}$ are functions of $\tilde{\mathbf{B}}\cdot\mathbf{B}$. Using, the result of [16], quoted in the introduction, the $\tilde{\mathbf{B}}$ maximising the expectation of any such quality factor is $\tilde{\mathbf{B}}^{opt} = \frac{1}{\gamma}\left\langle \mathbf{B} \right\rangle_{\mathbf{B}}$, where γ is a normalisation factor. As in [15], this may be re-written as $\lim_{k\to\infty} \frac{1}{\gamma k}\sum_{\tau=1}^{k} \mathbf{B}^\tau$, where the $\{\mathbf{B}^\tau\}$, $\tau = 1, \ldots, k$, are points randomly drawn from $P(\mathbf{B}/\{\boldsymbol{\xi}^\mu\})$. Using the results of the last paragraph, it may be shown in a couple of lines that $R^{opt} \equiv \tilde{\mathbf{B}}^{opt} \cdot \mathbf{B}$ is equal to $\sqrt{R^G}$. Thus, we have in principle a method for supervised learning, and using the replica method we have exactly worked out how well we can do.

The results are plotted in Figure 1 for $\rho = 1.0$ and Figure 2 for $\rho = 2.0$. On both figures, line 1 is the maximal variance strategy of [12], line 2 is Gibbs learning, which is always worse, and line 3 is optimal learning, which is always slightly better.

Expanding (16) in powers of small R^G and $\tilde{R}$, gives $\tilde{R} = \alpha\rho^4 R^G - \mathcal{O}(R^{G^3})$. Inserting this is (15) gives $R^G = \alpha\rho^4 R^G(1 - R^G) + \mathcal{O}(R^{G^3})$, which only has solutions for $\alpha > \alpha_c = \rho^{-4}$, the same critical α found in [12]. For $\alpha < \alpha_c$ is is *impossible* to use the training set to construct a $\tilde{\mathbf{B}}$ with a macroscopic overlap with $\mathbf{B}$. Later in this section we discuss how general this result is.

As mentioned in Section 2, it is possible to imagine a slightly larger hypothesis space in which the value of ρ itself is considered as a variable to be inferred. For $\alpha > \alpha_c$ the value of ρ can be deduced with an accuracy of order $1/\sqrt{N}$ from another saddle point equation, which amounts in fact to the condition that $R^G = q$. For $\alpha < \alpha_c$, however, it is *impossible* using the data to increase our knowledge about ρ, except in as much as we can infer that $\rho < \alpha^{-1/4}$.

The smallness of the improvement of optimal learning over the maximal variance algorithm demonstrates that for examples taken from distribution 1 there is no point in searching for a better algorithm than the maximal variance one. As α becomes large, R^{opt}, like R^{mv} tends to 1 as $1 - R^{opt} \sim 1/\alpha$. The ratio of the asymptotic coefficient is plotted against ρ on Figure 3. It has a maximum value of 1.17 at $\rho = 1.78$, so asymptotically optimal learning is at most 17% better in the high α limit.

3.3 Optimal learning for distribution 2

The situation is quite different, however, if the patterns come from distribution 2. A single pattern, say $\boldsymbol{\xi}^1$, gives the magnitude of the $\{B_i\}$, so that inference reduces to estimating the values of $\{s_i \equiv sgn(B_i)\}$. Hypothesis space is discrete.

Nevertheless, the calculation is substantially similar. Equation (14) is only modified in that the "entropy" term $G_2(R^G, \tilde{R})$ is replaced by another entropic function, $G_3(R^G, \tilde{R})$. Thus, the natural order parameters remain R^G and $\tilde{R}$. Since the disorder on component i is uncorrelated with the magnitude of B_i, R^G can also be interpreted as $\frac{1}{N}\sum_i s_i \tilde{s}_i$, where $\{\tilde{s}_i\}$ is hypothesis about $\{s_i\}$ drawn from the posterior probability of $\{s_i\}$ given $\{\boldsymbol{\xi}^\mu\}$. Eqn (16) remains the same, but eqn. (15) must be revised to

$$R^G = \int Dz \tanh(\tilde{R} + \sqrt{\tilde{R}}z) \tag{17}$$

R^G is plotted as line 4 in Figures 1 and 2. R^{opt}, which is again $\sqrt{R^G}$, is shown as line 5. For this distribution of examples, Gibbs learning soon overtakes the maximal variance strategy, and optimal learning is even better. In fact, as $\alpha \to \infty$, R^{opt} converges to 1 as

$$1 - R^{opt} = \frac{1}{\sqrt{2\pi\alpha\rho^2}} \exp\left(-\frac{\alpha\rho^2}{2}\right). \tag{18}$$

That is, exponentially quickly. An analogous result occurs in the supervised learning of a **B** with quantized components using a finite temperature stochastic dynamics [18]. Thus, optimal learning shows that there is distinct room for improvement over maximal variance learning. In principle, optimal learning can be implemented directly, as in [15], though this may be slow in practice.

Interestingly, the critical value, α_c, is once again $1/\rho^4$. This may be shown by expanding (17), and inserting the expansion of (16) given above. Thus the difference in the entropic terms caused by discreteness of the space turns out to be unimportant in the limit of low order, as might be expected.

3.4 Retarded Classification

Hansel, Mato and Meunier [19] recently studied the problem of supervised learning problem with a *parity machine* of a rule of the same form [1]. They demonstrated that in this problem a certain critical number of examples is required before any learning can occur, and called this behaviour *retarded generalisation.* They claim it will occur in all problems in which the underlying rule possesses certain discrete symmetries, an effect reminiscent of Landau theory.

By analogy, we call the related effect in unsupervised learning *retarded classification.* A stong plausibility argument suggests that if the clusters are of equal weight and related to each other by a reflection or rotational symmetry about an axis through the origin, then no direction perpendicular to this plane of symmetry or axis can be learnt without retarded classification using examples whose overlap with the centres of the clusters is only of order $1/\sqrt{N}$. This is simply because whenever there is such a symmetry the hypothesis space is described by a partition function with two terms: an energy term weighted by α which is *even* in the parameters which represent alignment in these directions, and an entropy term discouraging alignment which is also even. States for which these order parameters are zero are thus always a stationary points of the free energy, and so for low α, when the entropic term dominates, the minimal free energy is always for such states.

An example of such a distribution is sketched schematically in Fig 4. The irregular clusters are related by a rotational three–fold symmetry about axis parallel to vector $\mathbf{x}$. Although it is possible to begin to learn $\mathbf{x}$ immediately, no perpendicular direction (such as $\mathbf{y}_1$) can be learnt at all until a finite number of examples is presented. Retarded classification will occur: up to a certain α, it is impossible to recognise structure in the data (except the correlation in the direction $\mathbf{x}$ common to all clusters).

3.5 Neural networks

One network which might easily be used to learn this problem is a *perceptron*. The function this performs on an input $\boldsymbol{\xi}$ may be written $N(\boldsymbol{\xi}) = sgn(\mathbf{J}\cdot\boldsymbol{\xi})$, for some normalised N-vector $\mathbf{J}$. If $\mathbf{J}$ and $\mathbf{B}$ have overlap m, and $\boldsymbol{\xi}$ is taken from a distribution which gives (10), then $y \equiv \mathbf{J}\cdot\boldsymbol{\xi}/\sqrt{N}$ is distributed as

$$P(\mathbf{J}\cdot\boldsymbol{\xi}/\sqrt{N} = y) = \frac{1}{2}\sum_{\sigma=\pm1}\frac{1}{\sqrt{2\pi}}\exp(-\frac{(y-\sigma m\rho)^2}{2}), \tag{19}$$

so that D defined by (9) is $|H(\rho\mathbf{B}\cdot\mathbf{J})|$, where $H(x) \equiv \int_{-\infty}^{x} Dz$. This is a function just of $\mathbf{J}\cdot\mathbf{B}$, thus the best $\mathbf{J}$ is $\tilde{\mathbf{B}}^{opt}$ defined above. The best network which can be built has a quality $D = H(\rho R^{opt})$.

3.6 Deduction of the $\{\sigma^\mu\}$

Lastly we consider how well the $\{\sigma^\mu\}$ can be deduced. By inserting auxiliary field into the calculation of $< \ln Z >$ in Appendix 2, it can be shown that for either distribution of patterns, a $\{\tilde{\sigma}^\mu\}$ taken from $P(\{\sigma^\mu/\{\boldsymbol{\xi}^\mu\})$ (eqn. 4). has an overlap with the true $\{\sigma^\mu\}$ of

$$t^G \equiv \frac{1}{p}\sum_\mu \tilde{\sigma}^\mu\sigma^\mu = \int Dz \tanh\left[\rho^2 R^G + \sqrt{\rho}\sqrt{R}z\right], \tag{20}$$

while the optimal guess, $\{\sigma^\mu\}^{opt}$ has the higher value of

$$t^{opt} \equiv \frac{1}{p}\sum_\mu \tilde{\sigma}^{\mu,opt}\sigma^\mu = \int Dz \left|\tanh\left[\rho^2 R^G + \sqrt{\rho}\sqrt{R}z\right]\right|. \tag{21}$$

As α rises, and the clusters become better defined, t^G and t^{opt} rise. However, even in the $\alpha\to\infty$ limit neither tends to 1, so it is impossible to deduce with certainty from which cluster the data in the training set really comes. This is, of course, because the clusters overlap. Otherwise eqn. (4) would tend to 1 for one $\{\sigma^\mu\}$ and zero otherwise. The results for t^G and t^{opt} in the $\alpha\to\infty$ limit are the same for distributions 1 and 2, and are plotted in Fig. 5.

3.7 The effect of incorrect assumption on the hypothesis space

The preceding calculations can be generalized to the case where the value of ρ is not correctly guessed: one computes the performance of the model for a given value $\tilde{\rho}$. In that case R^G is not identical to q, and the quality of classification is now measured by $g \equiv \frac{R^G}{\sqrt{q}}$. We give here only the main results. Retarded classification occurs for any value of $\tilde{\rho}$. For $\tilde{\rho} > \rho$, the critical ratio α_c is equal to $1/\tilde{\rho}^2\rho^2$. This is a first order transition: g jumps from 0 to a finite value. For $\tilde{\rho} < \rho$, there is two transitions. Classification occurs for $\alpha > \alpha_c = 1/\rho^4$, with a continuous transition as for $\tilde{\rho} = \rho$. But there is also a spin-glass type phase for $\alpha_1 < \alpha < \alpha_c$. In this regime, $R^G = 0$ but q is finite. α_1 grows from 0 to $1/\rho^4$ when $\tilde{\rho}$ decreases from infinity to ρ.

Hence if the expected separation between the clusters is too large, more training examples are needed before generalization occurs, but there is a finite gain in generalization once the critical number is reached. If on the countrary the assumed separation is too small, there is first a kind of confusion phase, where the $\tilde{B}$'s become oriented but not correlated with the true B.

4 Multi–directional Models

It would be straightforward to generalise the problem of section 3 to $K > 1$. Unfortunately, a mathematical analysis of the result turns out to be closely related to that of supervised learning with "fully-connected committee machine", studied by Schwarze and Hertz [20]. In that problem the replica analysis is difficult to perform, and anyway it is thought very likely that replica symmetry is broken. While qualitative insight can certainly be obtained from such calculations, they cannot provide a quantitative comparison of algorithms, which is the purpose of this paper. We will therefore study a multi-directional problem in a simple and exactly solvable limit.

Examples are again normalised so that $\boldsymbol{\xi} \cdot \boldsymbol{\xi} = N$. A fraction $(1-\epsilon)/2$ have an overlap of value mN with an N-vector $\mathbf{B}^1$, and the same fraction have the same overlap with $\mathbf{B}^2$. The remaining ϵ are totally random, "noise examples". Note that in this problem examples have an extensive overlap with the centre of the cluster, so only of order 1 are needed. This in turn means that the problem may be analyses without replicas: the vectors of the training set span a space of much smaller dimensionality than N, so all are effectively perpendicular in the space perpendicular to that spanned by $\mathbf{B}^1$ and $\mathbf{B}_2$. All vectors of the same cluster are equivalent. We will allow the clusters to be correlated such that $\mathbf{B}^1 \cdot \mathbf{B}^2 = M$.

Our hypothesis about $\mathbf{B}^1$ and $\mathbf{B}^2$ is the pair of vectors $(\tilde{\mathbf{B}}^1, \tilde{\mathbf{B}}^2)$. The simple algorithm we will study is to choose $\tilde{\mathbf{B}}^1$ and $\tilde{\mathbf{B}}^2$ to minimise an energy,

$$\begin{aligned} E(\tilde{\mathbf{B}}^1, \tilde{\mathbf{B}}^2, \kappa) = \ & \textstyle\sum_\mu \ \Big(h\Theta(\kappa - \boldsymbol{\xi}^\mu \cdot \tilde{\mathbf{B}}^1)\Theta(\kappa - \boldsymbol{\xi}^\mu \cdot \tilde{\mathbf{B}}^2) \\ & - \ \boldsymbol{\xi}^\mu \cdot \tilde{\mathbf{B}}^1 \Theta(\boldsymbol{\xi}^\mu \cdot \tilde{\mathbf{B}}^1 - \kappa) - \boldsymbol{\xi}^\mu \cdot \tilde{\mathbf{B}}^2 \Theta(\boldsymbol{\xi}^\mu \cdot \tilde{\mathbf{B}}^2 - \kappa) \Big), \end{aligned} \tag{22}$$

where h is an arbitrary but very high number. The first term is a high penalty for any example with an overlap less than κ with both $\tilde{\mathbf{B}}^l$ vectors. The other two terms encourage the $\tilde{\mathbf{B}}^l$ vectors to be aligned to as many examples as possible with which they have an overlap greater that κ. For κ zero, the consequent $\tilde{\mathbf{B}}^1$ and $\tilde{\mathbf{B}}_2$ will be identical: the normalised sum of the training set; as κ rises, each $\tilde{\mathbf{B}}^l$ will have to decide which examples to remain aligned with , and eventually the examples are partitioned between $\tilde{\mathbf{B}}^1$ and $\tilde{\mathbf{B}}^2$; finally κ will reach a critical value, κ_c, when it is impossible to find $\tilde{\mathbf{B}}^1$ and $\tilde{\mathbf{B}}^2$ such that all the examples have an overlap greater than κ with one of them, and suddenly the first term of the energy will dominate.

This algorithm was chosen because of its similarity to a numerical one proposed by [11], which seems to work well in simple cases. Its other virtue is that it may be solved exactly. Once a $\tilde{\mathbf{B}}^l$ has chosen which examples it will be within κ of, it may be easily calculated. Therefore, since all examples in the same cluster are equivalent, the energy may be written as a function of six variables: n_1^1, n_2^1, $n^2, 1$, n_2^2, $\tilde{n}_1$ and $\tilde{n}_2$, where each n_l^m means how many of the examples of the mth cluster have overlap greater than κ with $\tilde{\mathbf{B}}^l$, and $\tilde{n}_l$ means how many of the noise examples have overlap greater than κ with $\mathbf{B}^l$. We will additionally assume the symmetries $n_1^1 = n_2^2$, $n_2^1 = n_1^2$ and $\tilde{n}_1$=$\tilde{n}_2$. The result is an energy in three variables which can be minimised by a search in the 3–dimensional space. We additionally assume, without loss of generality, that $R_1^1 \geq R_2^1$.

We define R^a as $\mathbf{B}^1 \cdot \tilde{\mathbf{B}}^1$, and R^b as $\mathbf{B}^1 \cdot \tilde{\mathbf{B}}^2$. The values of these parameters are plotted against κ on Fig. 6 for $p = 20$, $\epsilon = 0$, $m = .2$, $M = .1$. Fig. 7 shows the energy as a function of κ. For $\kappa < .261$ nothing happens. At $\kappa = .261$ each $\tilde{\mathbf{B}}^l$ is forced to give up one of the examples of the other cluster. At $\kappa = .347$ each $\tilde{\mathbf{B}}^l$ is composed only of the examples of one cluster, and its share of the noisy examples. κ_c is .369, which is marked by a vertical dashed line.

Of the variables plotted above only the energy variable is directly measurable. Its form, with regimes of stability and rapid change could itself be taken as evidence from structure in the data.

A curious feature of the curves in Fig. 6 are their "oscillations". The reason for this is that in order to make $\tilde{\mathbf{B}}^1$ and $\tilde{\mathbf{B}}^2$ have an overlap greater than κ with as many vectors as possible, $\tilde{\mathbf{B}}^l$ is overdominated by the outlying ones. When, as κ rises, an example is eventually relinquished, $\tilde{\mathbf{B}}^l$ snaps back towards the others. Note that for other realisations of the parameters, the behaviour may be qualitatively different in several ways. For example, for some the $\tilde{B}^l$ realease several of their component vectors simultaneously as κ rises slightly. For m small, the correct scaling of p is as $p \sim 1/m^2$. We introduce the rescaled variable $\tilde{p} \equiv pm^2$. Taking this limit gives a continuous version of the curves, in which the oscillations disappear.

Note that because of the different scaling in this problem, the two clouds of examples do not overlap at all. Therefore, retarded classification is not visible on the scale of $\tilde{p}$.

For $\kappa > \kappa_c$ the energy is dominated by the penalty term. If $\epsilon > 0$ the value of R^a initially rises as κ increases to climb, because the noise examples are discarded, but then falls as the true examples of the cluster are lost. If $\epsilon = 0$, R^a steadily as κ rises above κ_c. It seems that the $\kappa = \kappa_c$ is a good one at which to stop the algorithm. The values of R^a obtained by doing so for $M = .4$ are shown against $\tilde{p}$ in Fig. 8, for ϵ values of 0.0, 0.25

and 0.5.

The same features which make the above algorithm easy to analyse, make an analysis of optimal learning trivial. Given the form of the distribution, it is easy to deduce which examples are from which cluster, and which are noise. Because of the symmetry in the examples, the best guess for $\mathbf{B}^1$ can certainly be written as

$$\tilde{\mathbf{B}}^{1,opt} = \frac{1}{\gamma}\left(\sum_{\mu} \boldsymbol{\xi}^{\mu} + b\sum^{\nu} \boldsymbol{\eta}^{\nu}\right), \tag{23}$$

where γ is a normalisation constant, b is a number and we have used μ to label the examples from the first cluster and ν for the examples from the second. b can be found as the value which maximises $R^{a,opt} \equiv \tilde{\mathbf{B}}^{1,opt} \cdot \mathbf{B}^1$. On Fig. 8, $R^{1.opt}$ is plotted as the dashed lines; its improvement over the simple algorithm is, as expected, greater when the proportion of noise examples is higher.

We feel that although this model is simple, variations on it provide the best workshop in which to investigate theoretically unsupervised learning of several directions. The calculation may also easily be framed in terms of thermodynamics, so that even variations which are less transparent should be straightforward to analyse. The first pattern is how to alter the energy (22) so as to make is less susceptible to the noise in the examples. One would also like to know how the algorithms should be adjusted to model clusters of more complicated shape. Answering these patterns could be the task of another interesting paper.

5 Conclusion

We have introduced an inferential approach to unsupervised learning which has allowed us to define the optimal way in which it may be performed. We have shown in a simple problem that it is *impossible* to detect a structure in the data until a certain number of examples have been presented – an effect due to symmetry in the unknown underlying distribution. Thereafter the improvement of optimal learning over other techniques depends critically on the distribution of the examples.

We have also presented a simple model with more than one direction to be learnt, and studied an algorithm related to one used in practice. A great deal of work remains to be done on the selection of a good energy function, and our simple solvable model seems to provide a suitable framework within which to explore further, realistic variations.

Acknowledgements

One of us (T.L.H.W.) is grateful to the Laboratoire de Physique Statistique at the Ecole Normale Supérieure, Paris, for their hospitality while part of this work was carried out, and to St. John's College, Cambridge for full financial support.

References

[1] T.L.H. Watkin, A. Rau and M. Biehl, "The Statistical Mechanics of Learning a Rule", Rev. Mod. Phys. **65**, 499 (1993)

[2] Kohonen T.O. *Self-organization and associative memory.* Springer, Berlin, 1984.

[3] Bialek W., editor. *Princeton Lectures on Biophysics.* World Scientific, 1992.

[4] Linsker R. Self-organization in a perceptual network. *Computer*, 21:105–17, 1988.

[5] Atick J. J. Could information theory provide an ecological theory of sensory processing. *NETWORK*, 3:213–251, 1992.

[6] Nadal J.-P. and Parga N. Information processing by a perceptron in an unsupervised learning task. *NETWORK*, 4:295–312, 1993.

[7] Oja E. Neural networks, principal components, and subspaces. *Int. Journ. of Neur. Syst.*, 1:61–68, 1989.

[8] Hertz J., Krogh A., and Palmer R. G. *Introduction to the Theory of Neural Computation.* Addison-Wesley, Cambridge MA, 1990.

[9] Buhmann J. and Kuhnel H. Complexity optimized data clustering by competitive neural networks. *Neural Comp.*, 5:75–88, 1993.

[10] Geszti T. and Csabai I. Habituation in learning vector quantization. *Complex Systems*, 6:179–191, 1992.

[11] K. Rose, E. Gurewitz and G. C. Fox, *Phys. Rev. Letts.*, **65**, 945 (1990)

[12] M. Biehl and A. Mietzner, "Statistical Mechanics of Unsupervised Learning", Preprint Julius-Maximilians-Universität, Würzburg.

[13] S. Kullback, *Information Theory and Statistics*, (John Wiley, New-York, 1959)

[14] R. E. Blahut, *Principles and Practice of Information Theory* (Addison-Wesley, Cambridge MA, 1988)

[15] T.L.H Watkin, Europhys. Letts. **21**, 871 (1993)

[16] T.L.H. Watkin, R. Serneels and G.-J. Bex, *unpublished.*

[17] M. Opper, W. Kinzel, J. Kleinz, R. Nehl, *J.Phys.* **A 23** L581 (1990

[18] H.S. Seung, H. Sompolinsky and N. Tishby, *Phys. Rev. A.*,**45**, 6056 (1992)

[19] D. Hansel, G. Mato, C. Meunier. *Europhys. Letts* **20**, 471 (1992)

[20] H. Schwarze and J. Hertz, "Generalization in a fully–connected committee machine", Nordita Preprint No. 92/61S

Appendix 1

For completedness in this appendix we give the proof of the result quoted in Section (2.3), which is a generalisation by one of us (T.L.H.W.) of a proof contained in ref. [16]. Suppose that $\mathcal{Q}(\{\tilde{\mathbf{B}}^l\}, \{\mathbf{B}^l\})$ may be written as $\sum_l f_l(\tilde{\mathbf{B}}^l \cdot \mathbf{B}^l)$ for some smooth, increasing set of functions $\{f_l(x)\}$. Then

$$\left\langle Q(\{\tilde{\mathbf{B}}^l\}, \{\mathbf{B}^l\}) \right\rangle_{\{\mathbf{B}^l\}} = \lim_{k \to \infty} \frac{1}{k} \sum_{\tau=1}^{k} \sum_{l=1}^{K} f_l(\tilde{\mathbf{B}}^l \cdot \mathbf{B}^{l,\tau}), \tag{24}$$

where $\{\mathbf{B}^l\}^\tau$, $\tau = 1, \ldots, k$ are k *samplers* [15], sets of hypotheses drawn randomly from (2). Expanding the lth term gives

$$\frac{1}{k} \sum_{\tau=1}^{k} f_l(\tilde{\mathbf{B}}^l \cdot \mathbf{B}^{l,\tau}) = f\left(\tilde{\mathbf{B}}^l \cdot \left\langle \mathbf{B}^l \right\rangle\right) + \frac{1}{k} \sum_{\tau=1}^{k} \sum_{p=1}^{\infty} \frac{1}{p!} \frac{d^p}{dx^p} f_l(x) \bigg|_{x=\tilde{\mathbf{B}}^l \cdot <\mathbf{B}^l>} \left(\tilde{\mathbf{B}}^l \cdot \left(\left\langle \mathbf{B}^l \right\rangle - \mathbf{B}^{l,\tau}\right)\right)^p. \tag{25}$$

where $\left\langle \mathbf{B}^l \right\rangle \equiv \frac{1}{k} \sum_\tau \mathbf{B}^{l,\tau}$. But, if $\mathbf{B}^{l,\tau} \cdot \mathbf{B}^{l,\tau'} = q_l$ for some q_l and all $\tau \neq \tau'$, then $\left(\left\langle \mathbf{B}^l \right\rangle - \mathbf{B}^{l,\tau}\right) \cdot \left(\left\langle \mathbf{B}^l \right\rangle - \mathbf{B}^{l,\tau'}\right)$ equals $(q_l - 1)/k$, which goes to zero for k large. Therefore, for any $\tilde{\mathbf{B}}^l$, $\tilde{\mathbf{B}}^l \cdot \left(\left\langle \mathbf{B}^l \right\rangle - \mathbf{B}^{l,\tau}\right)$ can be of order 1 for at most of order 1 of the samplers, and of order $1/\sqrt{k}$ for the rest. In either case all the terms of the sum over p disappear. Thus eqn. (24) gives

$$\left\langle Q(\{\tilde{\mathbf{B}}^l\}, \{\mathbf{B}^l\}) \right\rangle_{\{\mathbf{B}^l\}} = \sum_{l=1}^{K} f_l\left(\tilde{\mathbf{B}}^l \cdot \left\langle \mathbf{B}^l \right\rangle\right). \tag{26}$$

Therefore, since the $\{f_l(x)\}$ are increasing functions, the optimal $\tilde{\mathbf{B}}^l$ is $\tilde{\mathbf{B}}^l = \left\langle \mathbf{B}^l \right\rangle$, *independent* of the quality measure $\mathcal{Q}$.

Note that the above argument is accurate to order $1/\sqrt{k}$, and for $k << N$. Eqn. (26) is therefore accurate to leading order in N for N large.

Appendix 2

The method of replicas uses the identity $\ln Z = \lim_{n \to 0} \frac{1}{n}(Z^n - 1)$. Inserting (11) into the definition of Z, eqn. (3), gives

$$Z^n \propto \prod_{\alpha=1}^{n} \int d\mathbf{B}^\alpha \delta(\mathbf{B}^\alpha \cdot \mathbf{B}^\alpha - 1) Tr_{\{\sigma^\mu\}} \prod_\mu \exp\left(-\sqrt{N} \boldsymbol{\xi}^\mu \mathbf{B}^\alpha \rho \sigma^\mu - \boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\mu / 2\right). \tag{27}$$

We now perform the average of Z^n over the possibilities for $\{\boldsymbol{\xi}^\mu\}$, simplify the expression by introducing order paramenter, and take the limit of $n \to 0$. The values of the four order parameters R^G, $\tilde{R}$, q and $\tilde{q}$ are, as usual, those which maximise it,

$$\langle \ln Z \rangle = \max_{R^G, \tilde{R}, q, \tilde{q}} [\alpha G_1(R^G, q) + G_2(R^G, \tilde{R}, q, \tilde{q})], \tag{28}$$

where

$$
\begin{aligned}
G_1 &= (1-q)\rho^2/2 + \int Dz \, \ln\cosh\left[R^G\rho^2 + \rho\sqrt{q}z\right] \\
G_2 &= \frac{1}{2(1-q)} - \tilde{R}R^G - \tilde{R}^2(1-q)/2 + \frac{1}{2}\ln(1-q)
\end{aligned}
\tag{29}
$$

The right hand side of the term in the square brackets in (28) is invariant under $R^G \to -R^G$, and $\tilde{R} \to -\tilde{R}$. The saddle point equations for R^G, $\tilde{R}$, q and $\tilde{q}$, are

$$
\begin{aligned}
\tilde{R} &= \alpha\rho^2 \int Dz \tanh\left[R^G\rho^2 + \rho\sqrt{q}z\right] \\
\tilde{q} &= \alpha\rho^2 \int Dz \tanh^2\left[R^G\rho^2 + \rho\sqrt{q}z\right] \\
R^G &= \tilde{R}(1-q) \\
q &= (\tilde{q} + \tilde{R}^2)(1-q)^2.
\end{aligned}
$$

Using the curious identity proved in Appendix 3, we obtain that $\tilde{R} = \tilde{q}$ and $R^G = q$, so that $\langle \ln Z \rangle$ can be written in the form of eqn (14), with G_1 as a function of R^G and G_2 as a function of $\tilde{R}$ and R^G.

For distribution 2, Z^n becomes

$$
Z^n = \underset{\{\sigma_\alpha^\mu\}}{Tr} \; \underset{\{s_i^\alpha\}}{Tr} \prod_{\mu,i} \frac{1}{2} \exp\left(\frac{\rho}{\sqrt{N}} \sum_\alpha \xi_i^\mu \sigma_\alpha^\mu s_i^\alpha\right) \tag{30}
$$

Eventually much the same algebra leads to G_2 being replaced by

$$
G_3 = -\tilde{R}R^G + \tilde{q}(1-q)/2 + \int Dz \ln\cosh(\tilde{R} + \sqrt{q}z). \tag{31}
$$

Again we obtain $\tilde{R} = \tilde{q}$ and $R^G = q$, and easily obtain (17).

Appendix 3

Here we prove the identity that for any a,

$$
\int Dt \tanh(at + a^2) = \int Dt \tanh^2(at + a^2). \tag{32}
$$

We begin by noting that

$$
\int Dt \frac{d}{dt} \tanh(at + a^2) = a \int Dt \left(1 - \tanh^2(at + a^2)\right) = \int Dt \, t \tanh(at + a^2), \tag{33}
$$

where the last term follows from the first by integration by parts. Therefore,

$$
\int Dt \left[\tanh(at + a^2) - \tanh^2(at + a^2)\right] = \frac{1}{a^2} \int (at + a^2) \tanh(at + a^2) Dt - 1. \tag{34}
$$

Let $u = at + a^2$. Then (34) equals

$$\frac{1}{\sqrt{2\pi}a^3} \int u \tanh(u) e^{-(u-a^2)^2/(2a^2)} \mathrm{d}u - 1, \tag{35}$$

which, if the exponentials of the integrand are rearranged, becomes

$$\frac{1}{\sqrt{2\pi}a^3} \int \left[u e^{-(u-a^2)^2/(2a^2)} - \frac{u e^{-(u^2+a^4)/(2a^2)}}{\cosh u} \right] \mathrm{d}u - 1. \tag{36}$$

The second term in the square bracket is odd, and hence disappears in the integral. The rest can be easily performed, to show that (36) is zero, which proves the identity.

Figure Captions

- Fig. 1 Unsupervised learning for $\rho = 1.0$. Line 1 is the maximal variance algorithm. Line 2 and line 3 are Gibbs and optimal learning respectively, if patterns are taken from distribution 1. Lines 4 and 5 are Gibbs and optimal learning respectively, if the patterns are from distribution 2.

- Fig. 2 Shows the same quantities as Fig. 1, but for $\rho = 2.0$

- Fig. 3 The ratio of the coefficients of the α^{-1} decay for optimal and maximal variance learning as $\alpha \to \infty$ and for patterns from distribution 1. The height of the curve is a measure of the advantage of the optimal strategy.

- Fig 4. A sketch of a distribution of patterns such that there will be retarded classification.

- Fig 5. The extent to which it is possible to ascertain from which cluster patterns are drawn in the $\alpha \to \infty$ limit. Lines 1 and 2 show respectively t^G and t^{opt}, which are defined in the text. The curve is the same for distributions 1 and 2.

- Fig 6. The values of R^a (line 1) and R^b (line 2) against κ, for the simple algorithm of section 2.

- Fig 7. The value of energy against κ.

- Fig 8. Learning by the simple algorithm (solid lines), and by optimal learning (dashed line) for $M = 0.4$ and ϵ values of 0.0, 0.25 and 0.5.